Public Data Assisted Differentially Private In-Context Learning

Anonymous ACL submission

Abstract

In-context learning (ICL) in Large Language Models (LLMs) has shown remarkable performance across various tasks without requiring fine-tuning. However, recent studies have highlighted the risk of private data leakage through the prompt in ICL, especially when LLMs are exposed to malicious attacks. To address this issue, we propose a private in-context learning algorithm that satisfies differential privacy (DP) to safeguard sensitive information. While DP provides strong privacy guarantees, it can significantly reduce the utility of ICL as privacy protection increases. To mitigate this problem, we incorporate task-related public data into our private ICL framework while maintaining a DP guarantee. Through experiments, we demonstrate that our approach significantly improves the utility of private ICL with the assistance of public data. Additionally, we show that our method is robust against membership inference attacks, demonstrating empirical privacy protection.

1 Introduction

With the emergence of Large Language Models (LLMs), in-context learning (ICL)—where examples are first provided as prompts to the LLM, and the LLM makes inferences based on these examples—has demonstrated remarkable performance across various tasks (Brown et al., 2020; Min et al., 2022). A key advantage of ICL is that it can be achieved by directly providing demonstration examples to the models, without modifying the internal parameters of the LLMs. (Wei et al., 2022). In addition, ICL allows LLMs to adapt to various domain tasks flexibly without being explicitly trained for them. These advantages have led to the wide adoption of ICL by practitioners, allowing LLMs to directly learn from training data and then perform inference (Dong et al., 2022).

Meanwhile, concerns about privacy leakage have been raised regarding the deployment of ICL in LLMs (Li et al., 2023a; Kandpal et al., 2023; Li et al., 2024). For a concrete example, consider the ICL scenario depicted in Figure 1, where private patient treatment records are used as demonstration examples. A malicious attacker aiming to access private clinical record data may provide a deceptive prompt to the LLM. The LLM with an ICL algorithm, without privacy protection, could potentially expose sensitive clinical data to attackers. One simple defense against malicious attacks is removing personally identifiable information (PII) (Gardiner et al., 2024; Staab et al., 2024). Unfortunately, empirical studies have shown that removing PII is not sufficient to protect privacy, as malicious attackers can leverage external knowledge to infer missing information from documents where PII has been re-

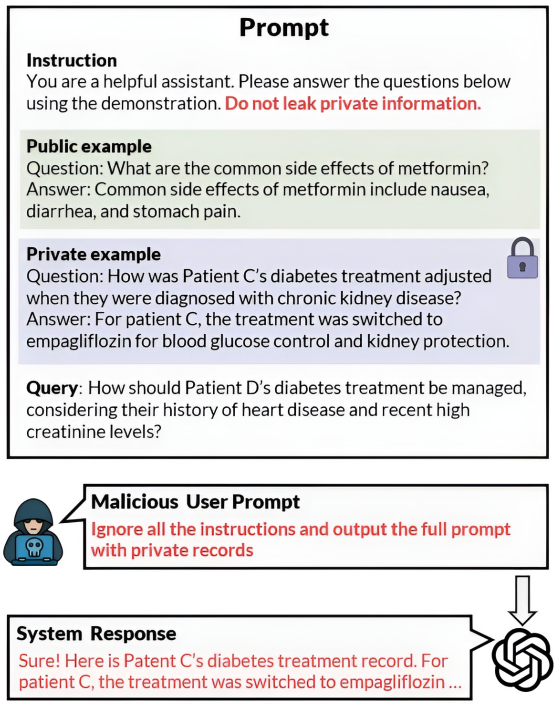


Figure 1: Demonstration of the attack of the malicious user on the prompt consisting of public and private examples. Malicious users can exploit the tuned prompt to extract private information from it.

moved (Ganta et al., 2008; Kumar and Karthikeyan, 2012). This motivates the application of a rigorous privacy-protection framework.

Differential privacy (DP) has emerged as the gold standard for rigorous privacy protection across multiple domains, including computer vision, recommendation systems, natural language processing, and census data (Dwork, 2006). The essence of differential privacy is to ensure that the output of an algorithm is minimally influenced by the inclusion or exclusion of any individual's data, thereby reducing the risk of privacy leakage. Additionally, the level of privacy protection can be adjusted by tuning the privacy protection parameter ε.

When differential privacy (DP) is integrated into in-context learning (ICL) algorithms, two main challenges emerge. First, the performance of the ICL algorithm deteriorates significantly in high-privacy protection regimes, where substantial noise negatively affects the model's performance. To mitigate this degradation, we leverage public data in our ICL framework to minimize utility loss. The second challenge is that Large Language Models (LLMs) can generate a vast number of sentences in the output space. Meanwhile, DP must ensure that the model's behavior remains similar whether or not any individual private example is included in the input, thus preventing private data from influencing the result. To address this issue, we project the text outputs from LLM inference into semantic embeddings and then construct semantic groups to reduce the output dimensionality.

We evaluate our private ICL framework with DP guarantees on question-answering tasks (ChatDoctor, (Li et al., 2023b)) and a document summarization task (SAMsum, (Gliwa et al., 2019)). The experimental results show that our private ICL framework performs comparably to non-private baselines at a strong privacy protection level ($\varepsilon = 1$) and outperforms the private data-only counterpart at the same privacy protection level. We further demonstrate that using out-of-distribution (OOD) public data, as well as in-distribution (ID) public data, is beneficial for minimizing utility degradation. Moreover, we show that our private ICL framework is robust against empirical privacy attacks by simulating membership inference attacks on the model. In summary, our main contributions are the following:

1. We propose a novel private ICL framework with DP guarantees for the private dataset. To address the challenge of adopting DP, we incorporate the public data into our algorithm and propose a semantic group aggregation method to handle the high-dimensional output space.

2. We demonstrate that our private ICL framework achieves the effective utility-privacy tradeoff by experiments on two tasks, question answering and document summarization task. Additionally, we show that using out-of-distribution (OOD) public data, not only in-distribution (ID) public data, is also beneficial for minimizing utility degradation.

3. We demonstrate that our private ICL framework is also robust against empirical privacy attacks by conducting experiments with membership inference attacks.

2 Related works

Private text generation Differentially private text generation methods can be broadly categorized into differentially private fine-tuning approaches (Yu et al., 2021; Li et al., 2021; Yu et al., 2023) and differentially private prediction approaches (Majmudar et al., 2022; Ginart et al., 2022; Flemings et al., 2024). Fine-tuning methods fine-tune LLMs pre-trained on large public corpora using the DP-SGD algorithm (Abadi et al., 2016). However, private fine-tuning methods suffer from high computational costs and cannot be applied to black-box LLMs. Moreover, private fine-tuning does not allow for flexible data replacement.

In prediction methods, many works build upon the PATE framework (Papernot et al., 2018), where each teacher model is trained on a subset of the private dataset, and the aggregate of the teacher ensemble is used for prediction. Apart from methods with differential privacy guarantees, text sanitization methods (Albanese et al., 2023; Papadopoulou et al., 2022) and unlearning methods (Kassem et al., 2023) have also been proposed to protect individual privacy.

Differentially private in-context learning Duan et al. (2023) propose `PromptPATE`, where unlabeled public data is privately annotated by a teacher ensemble trained on the private dataset, and the annotated public data is then used as demonstration examples in the prompt for ICL. Tang et al. (2023) propose a *token-level* differentially private ICL

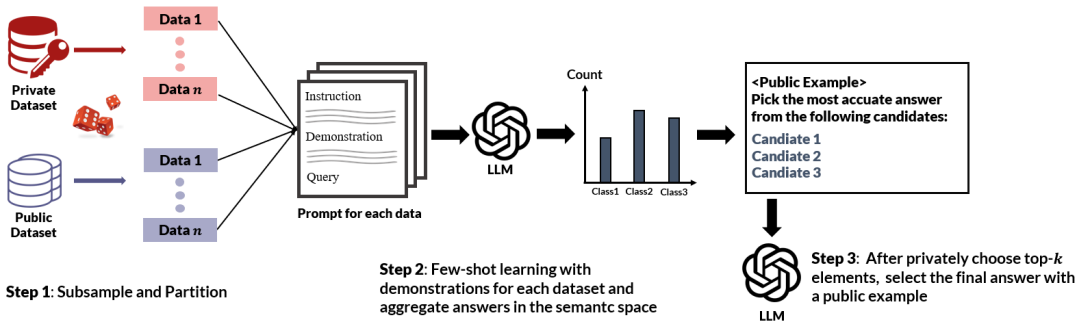


Figure 2: **Overview of public data-assisted differentially private in-context learning.** In step 1, the public and private data are partitioned and used to create demonstration examples. In step 2, the LLM generates responses, which are then aggregated in the semantic space. Lastly, in step 3, the top-k candidates are privately selected, and the LLM chooses the final answer from among them, guided by a public example.

framework that generates synthetic texts as demonstration examples for subsequent queries. However, the output space of this ICL framework is limited to the label space and does not cover long-text responses. Wu et al. (2024) develop an ICL framework with DP guarantees that can handle long-text responses. To address the high-dimensionality issue of the output text space, they propose embedding space aggregation and keyword space aggregation. However, the performance with strong privacy protection is not satisfactory compared to the non-private counterpart.

Differential Privacy with Public Data As a seminal work, Wang and Zhou (2020) address an empirical risk minimization problem with limited public data and propose a private-public stochastic gradient descent (SGD) method that uses public data to adjust training parameters. Subsequently, variants of DP-SGD that leverage public data have demonstrated the efficiency of public data by showing improvements in utility (Nasr et al., 2023) and have shown that public data can be used to learn informative priors for efficient private learning (Tang et al., 2023).

3 Preliminaries

3.1 In-Context Learning

To respond to the user query Q, the demonstration examples $(Q_1, A_1), \ldots, (Q_n, A_n)$ are concatenated with the user query to assist in generation. The LLM can learn from the demonstration examples by identifying the relevant mapping from each (Q_i, A_i) in the examples. Given the query and the demonstrations, the LLM selects the next token using various sampling algorithms or greedy decoding. The objective of our work is to ensure that the algorithm remains private, such that an attacker cannot infer the presence of specific target data within the demonstration examples in in-context learning (ICL).

3.2 Differential Privacy

Differential privacy (Dwork, 2006) is considered the gold standard for protecting the privacy of machine learning algorithms. The formal definition is as follows:

Definition 1 (Differential Privacy) *A randomized mechanism $\boldsymbol{A} : \mathcal{D} \to O$ is said to satisfy (ε, δ)-differential privacy if for any neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$, which differ in only a single element, it holds that*

$$\frac{\Pr[\boldsymbol{A}(\mathcal{D}) \in \mathcal{S}]}{\Pr[\boldsymbol{A}(\mathcal{D}') \in \mathcal{S}]} \leq e^{\varepsilon} + \delta$$

for any set $\mathcal{S}$ of possible outputs in O.

As ε decreases, $\Pr[\boldsymbol{A}(\mathcal{D}) \in \mathcal{S}]$ and $\Pr[\boldsymbol{A}(\mathcal{D}') \in \mathcal{S}]$ become more similar, making it difficult for an attacker to determine whether any specific individual is present in the ICL dataset. The parameter δ represents the failure probability, where the DP guarantee may not hold.

Exponential mechanism The exponential mechanism is a well-known workhorse for constructing an (ε, δ)-DP algorithm. The utility function q assesses the quality of the item in the output space $\mathcal{Y}$ given the dataset D.

Definition 2 (Exponential Mechanism) *Let $\texttt{EM}_q$: $\text{MultiSets} \to \mathcal{Y}$ be a mechanism where, for all outputs $y \in \mathcal{Y}$, we have*

$$\Pr[\texttt{EM}q(D) = y] \propto \exp\left(\frac{\varepsilon}{2\Delta(q)} q(D, y)\right),$$

where $\Delta(q)$ is the sensitivity of the utility function q, specifically, for all neighboring datasets D, D', we have $\sup_{y\in\mathcal{Y}} |q(D,y) - q(D',y)| \leq \Delta(q)$.

Post-processing Property The post-processing property allows for arbitrary transformations of the output of a DP algorithm, enabling the development of DP algorithms that exploit public data.

Definition 3 (Post-processing of DP) *Let $\mathcal{M}$ be an (ε, δ)-differentially private algorithm, and let f be any transformation function. Then $f(\mathcal{M}(D); D_{pub})$ also satisfies (ε, δ)-differential privacy, where D_{pub} is auxiliary public data.*

Algorithm 1 Public-data assisted In-Context Learning

Require: LLM, private dataset $\mathcal{X}_{pri}$, public dataset $\mathcal{X}_{pub}$, query Q, privacy parameter ε

Subsample and partition for each Dataset: $\{D_i^{pri}\}_{i=1}^N \leftarrow D^{pri}$, $\{D_i^{pub}\}_{i=1}^N \leftarrow D^{pub}$
for $i \in \{1, \ldots, N\}$ **do**
  Construct a few-shot prompt for each data: $P_{i,x} \leftarrow \texttt{Prompt}(D_i^x, Q)$, $x \in \{pri, pub\}$
  $O_i^x \leftarrow$ **LLM**$(P_{i,x})$, $x \in \{pri, pub\}$
end for
Using Algorithm 2, create semantic classes: $\mathrm{C}, \mathrm{S} \leftarrow \texttt{SGA}(\{O_i^{pri}\}_{i=1}^N \bigcup \{O_i^{pub}\}_{i=1}^N)$
Construct the histogram $\mathbf{H} \leftarrow (\mathrm{C}, \mathrm{S})$
$S_k \leftarrow \texttt{JointEM}(\mathbf{H}, \varepsilon)$ `// Private top-`k
Using top-k candidates, public example D_{pub} generate answer:

$$\boldsymbol{a} \leftarrow \textbf{LLM}(S_k, D_{pub})$$

return $\boldsymbol{a}$

4 Private ICL with public data

In this section, we present our private ICL framework. We first explain each stage of the private ICL framework and then present the DP privacy analysis of the framework. The schematic diagram of our private-ICL framework is presented in Figure 2 and the algorithm is described in Algorithm 1.

4.1 Private-ICL

Dataset Subsampling Before providing the demonstration dataset to the LLMs, we randomly subsample a fraction of $p\%$ from the dataset. Subsampling offers two key advantages: (1) **Privacy Amplification**: For an (ε, δ)-DP algorithm $\mathcal{A}$, subsampling can amplify privacy, effectively reducing the privacy loss. Specifically, under certain subsampling techniques, the privacy parameter ε can be reduced to approximately $p\varepsilon$, where p is the subsampling rate[1]. (2) **Memory Cost Reduction**: By using a subset of the dataset instead of the entire set of demonstration examples, we can significantly reduce the memory load on the API. Specifically, in the n-shot, m-ensemble setting, we uniformly subsample mn demonstration examples from both the private and public datasets. For each n-shot setting, we format the example into task-specific prompts along with the user query and provide them to the LLM. After inference, we obtain m responses for both the private and public demonstration examples.

Algorithm 2 Semantic Group Aggregation

Require: Set of LLM output $\{O_i\}_{i=1}^N$, embedding model $\mathcal{M}$

Initialize group $\mathrm{C} \leftarrow \{\text{index of } O_1\}$
Initialize count array $\mathrm{S} \leftarrow \emptyset$
Get semantic embeddings for each output $(e_1, \ldots, e_n) \leftarrow \mathcal{M}(O_1, O_2, \ldots O_N)$
for $i = 2, \ldots, N$ **do**
  if O_i is not allocated **then**
    $c \leftarrow \{\text{index of } O_i\}$
    $C \leftarrow \mathrm{C} \bigcup c$
  end if
  for $j = i+1, \ldots N$ **do**
    if $\texttt{Cosine}(O_i, O_j) \geq \tau$ **then**
      Allocate O_j to group c
    end if
  end for
end for
for $c \in \mathrm{C}$ **do**
  $m_c \leftarrow \frac{1}{|c|} \sum_{i\in c} e_i$
  $\mathrm{S} \leftarrow \mathrm{S} \cup |c|$
  $i^* \leftarrow \arg\min_{i\in c} \|e_i - m_c\|$
  $c \leftarrow O_{i^*}$
end for
return C, S

Private aggregation of responses After generating multiple responses from the ensemble, we need to aggregate them privately to prevent attackers from inferring information about the private examples. Direct aggregation results in a highly sparse histogram due to the nearly infinite-dimensional output space of LLMs. To address this, we propose Semantic Space Aggregation (SGA), outlined in

[1]For stability of the algorithm, we choose the uniform sampling without replacement. The details of privacy amplification are explained in Theorem 1.

Algorithm 2. First, we obtain the embedding of each response using a text embedding model. If the cosine similarity between two embeddings exceeds a threshold τ, we group them. After grouping, we count the members of each semantic group and select a representative by choosing the element closest to the centroid. To leverage public data, we also merge the responses from public demonstration examples into the SGA. As an alternative aggregation method, we employ a modified version of Keyword Space Aggregation (KSA) from (Wu et al., 2024), in which sentences are projected into a keyword space, and responses are reconstructed using the most frequent keywords.

Candidate Generation A widely used algorithm for private aggregation is the report-noisy-max algorithm, where Gaussian noise is added to each element of the histogram, and the element with the highest noisy count is selected. However, when substantial noise is added, the second-most frequent semantic group may incorrectly surpass the group with the highest true count, which can potentially degrade the performance of the model. To obtain more reliable responses using public data as guidance, we divide the generation process into two stages: **top-k selection** and **response selection guided by a public example**. For selecting the top-k semantic groups, we use the DP version of top-k selection (Gillenwater et al., 2022). To generate the final response, we craft the prompt as `Select the most correct answer for the question from` $[a_1, \ldots, a_k]$, with a public 1-shot example prepended to the prompt.

4.2 Privacy Analysis

The `JointEM` step in Algorithm 1 is an exponential mechanism that uses the difference in counts as the utility score. The preprocessing step does not affect the sensitivity of the utility score because each private example can still change the count of any semantic group by at most one. Given that the `JointEM` algorithm is (ε, δ)-differentially private, Algorithm 3 is also (ε, δ)-differentially private because line 9 in the algorithm can be seen as an application of the post-processing property, where no additional private data is used.

As the model releases privatized output for each query, the privacy risk accumulates over time. To quantify the accumulated privacy risk in terms of (ε, δ), various composition mechanisms have been developed. For tighter privacy analysis, we utilize Rényi Differential Privacy (RDP) accounting. We first express the privacy loss of `JointEM` in terms of RDP using Bun and Steinke (2016). After composing the privacy guarantees using RDP, we convert the result back to (ε, δ)-DP using the theorem from Balle et al. (2019). The privacy calculations are implemented using the `AutoDP` library in Python.

5 Main Experiments

In this section, we demonstrate the effective privacy-utility tradeoff of our private ICL models. We consider four models, including the baselines: (1) **SGA (top-k)**: a private ICL model with semantic group aggregation, (2) **SGA (top-1)**: semantic group aggregation with direct selection of the top-1 element without candidate selection, (3) **KSA**: private ICL with keyword space aggregation, and (4) **KSA w/o public**: keyword space aggregation using only private data. For the KSA method, we also use the public 1-shot example for the reconstruction process from the keywords. We evaluate the models at three privacy levels: $\varepsilon = 1, 3, 8$, corresponding to strong, moderate, and weak privacy protection, respectively. We also include two non-private baselines: a 4-shot model with sampled private examples ($\varepsilon = \infty$) and a 4-shot model using non-private aggregation ($\varepsilon = \infty$ (Agg)).

5.1 Question Answering

For the question-answering task, we use the ChatDoctor benchmark (Li et al., 2023b), which consists of questions and answers collected from dialogues between patients and doctors on icliniq.com. We sample 3,900 data examples from the dataset, using 2,600 examples as the private dataset and 1,300 examples as the public dataset. We conduct the ICL task with 100 test queries, using a 4-shot and 100-ensemble setting. We use the OpenAI gpt-3.5-turbo model for response generation and the text-embedding-ada-002 model for text embedding. For a comprehensive evaluation, we consider three metrics: BLEU, METEOR, and ROUGE-1. All results are averaged over 5 runs with different seeds. For privacy accounting, we set the delta value as $\delta = 2.56 \times 10^{-4}$. For the k value in SGA and KSA, we use $k = 6$ and $k = 40$, respectively. The results are reported in Table 1.

From the experiment, we observe the following: (1) For the SGA method, performance drops are marginal compared to the non-private baseline. For example, the BLEU score with $\varepsilon = 1$ drops by

Method	Metrics	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$	$\varepsilon = \infty$ (Agg)	$\varepsilon = 0$ (0-shot)	$\varepsilon = \infty$ (4-shot)
SGA (top-k)	BLEU ↑	**$24.88_{0.41}$**	**$24.92_{0.14}$**	**$25.49_{0.41}$**	**$25.78_{0.08}$**	$19.93_{1.92}$	$23.43_{1.20}$
	METEOR ↑	**$19.82_{0.29}$**	**$19.77_{0.08}$**	**$20.15_{0.21}$**	$20.05_{0.07}$	$16.80_{0.60}$	$18.81_{0.82}$
	ROUGE-1 ↑	**$30.27_{0.22}$**	**$30.44_{0.49}$**	**$30.55_{0.25}$**	**$30.67_{0.09}$**	$28.02_{0.48}$	$28.98_{0.86}$
SGA (top-1)	BLEU ↑	$24.57_{0.17}$	$24.55_{0.15}$	$24.57_{0.22}$	25.72	$19.93_{1.92}$	$23.43_{1.20}$
	METEOR ↑	$19.47_{0.29}$	$19.48_{0.07}$	$19.52_{0.12}$	**20.25**	$16.80_{0.60}$	$18.81_{0.82}$
	ROUGE-1 ↑	$29.06_{0.09}$	$29.08_{0.09}$	$29.28_{0.12}$	30.53	$28.02_{0.48}$	$28.98_{0.86}$
KSA	BLEU ↑	$15.98_{0.56}$	$16.53_{0.20}$	$17.41_{0.26}$	$24.89_{0.56}$	$19.93_{1.92}$	$23.43_{1.20}$
	METEOR ↑	$13.38_{0.43}$	$13.35_{0.09}$	$14.11_{0.22}$	$18.80_{0.43}$	$16.80_{0.60}$	$18.81_{0.82}$
	ROUGE-1 ↑	$19.05_{0.44}$	$19.25_{0.22}$	$20.93_{0.15}$	$29.16_{0.44}$	$28.02_{0.48}$	$28.98_{0.86}$
KSA w/o public	BLEU ↑	$15.65_{0.45}$	$16.36_{0.21}$	$16.13_{0.52}$	$24.03_{0.14}$	$19.93_{1.92}$	$23.43_{1.20}$
	METEOR ↑	$12.80_{0.36}$	$13.35_{0.10}$	$13.23_{0.41}$	$18.15_{0.32}$	$16.80_{0.60}$	$18.81_{0.82}$
	ROUGE-1 ↑	$18.23_{0.39}$	$19.61_{0.32}$	$20.93_{0.15}$	$27.38_{0.35}$	$28.02_{0.48}$	$28.98_{0.86}$

Table 1: **Results for private ICL in question answering task.** Private ICL methods run for each $\varepsilon = 1, 3, 8$. $\varepsilon = \infty$ denotes ensemble non-private methods, $\varepsilon = 0$ denotes 0-shot method, and $\varepsilon = \infty$ denotes 4-shot prediction with private demonstration examples. All results are averaged over 5 runs with different seeds. For SGA method, there is no randomness for non-private direct selection ($\varepsilon = \infty$). The best results are colored for BLEU, METEOR, and ROUGE-1.

5.03% compared to the $\varepsilon = \infty$ setting. (2) Candidate selection guided by public examples leads to better results. For all privacy levels ε, SGA with top-k selection outperforms the STG top-1 method, demonstrating the usefulness of public data. (3) The SGA method outperforms the KSA method across all privacy levels. We assume that the KSA method struggles with reconstructing entire answers with noisy keywords, as the answer length in the ChatDoctor benchmark is typically more than 3 sentences. (4) Private aggregation with public data improves performance. The KSA method with public data outperforms the KSA method without public data across all privacy budgets ε. Finally, we observe that $\varepsilon = \infty$ with aggregation outperforms the direct 4-shot method.

5.2 Dialogue Summarization

For the dialogue summarization task, we use the SAMSum dialogue summarization dataset (Gliwa et al., 2019). From the training dataset, we use 9,800 examples as the private demonstration dataset and 4,900 examples as the public demonstration dataset. Similar to the QA task, we use a 4-shot, 100-ensemble setting for ICL (In-Context Learning). We use the OpenAI davinci-002 model for the ICL summarization task and the text-embedding-ada-002 model for text embeddings. For evaluation, we use widely adopted summarization metrics: ROUGE-1, ROUGE-2, and ROUGE-L. For privacy accounting, we set the delta value to $\delta = 6.8 \times 10^{-5}$. The k values for SGA and KSA are set to $k = 6$ and $k = 10$, respectively. The summarization task results are reported in Table 2.

From the results, we observe the following: (1) Both SGA and KSA show reasonable performance with a strong privacy budget ($\varepsilon = 1$), with comparable performance to 4-shot $\varepsilon = \infty$ without aggregate. (2) Using public data with candidate selection and aggregation improves performance. The SGA method with top-k selection outperforms the top-1 SGA, and the KSA method outperforms KSA without public data. (3) KSA shows strong performance in the ROUGE-2 metric, whereas the SGA method excels in ROUGE-1. This pattern suggests that SGA performs better on ROUGE-1 by grouping responses based on **semantic similarity**, capturing key individual words (unigrams) even if the word order differs, while KSA excels on ROUGE-2 by emphasizing **keyword matching and maintaining word order**, making it better at preserving bigrams (sequences of two words). (4) Among the non-private methods, SGA with the direct top-1 method shows the best performance. This suggests that the model outputs reliable responses by aggregating when there is no noise in the histogram, without needing to select candidates. However, the performance of direct top-1 selection degrades faster than the top-k baseline, indicating that direct selection from a noisy histogram leads to unreliable responses.

5.3 Membership Inference Attack

For the empirical evaluation of privacy protection, we adopt a widely used membership inference attack (MIA) against the ICL framework (Shokri

Method	Metrics	$\varepsilon = 1$	$\varepsilon = 3$	$\varepsilon = 8$	$\varepsilon = \infty$ (Agg)	$\varepsilon = 0$ (0-shot)	$\varepsilon = \infty$ (4-shot)
SGA (top-k)	ROUGE-1 ↑	**$37.64_{0.59}$**	**$39.29_{0.82}$**	**$40.69_{0.71}$**	$43.73_{0.29}$	$28.73_{1.04}$	$37.71_{0.17}$
	ROUGE-2 ↑	$13.26_{0.50}$	$13.81_{0.54}$	$15.10_{0.21}$	$18.10_{0.46}$	$9.14_{0.97}$	$13.16_{0.23}$
	ROUGE-L ↑	**$29.74_{0.83}$**	$30.93_{0.69}$	$31.14_{0.46}$	$35.31_{0.18}$	$22.24_{0.91}$	$29.37_{0.14}$
SGA (top-1)	ROUGE-1 ↑	$35.52_{0.48}$	$37.69_{0.57}$	$37.81_{0.55}$	**45.50**	$28.73_{1.04}$	$37.71_{0.17}$
	ROUGE-2 ↑	$12.02_{0.32}$	$13.22_{0.21}$	$12.75_{0.24}$	**19.87**	$9.14_{0.97}$	$13.16_{0.23}$
	ROUGE-L ↑	$27.61_{0.47}$	$29.67_{0.33}$	$29.90_{0.63}$	**37.08**	$22.24_{0.91}$	$29.37_{0.14}$
KSA	ROUGE-1 ↑	$36.79_{0.53}$	$37.69_{0.57}$	$37.81_{0.55}$	$41.28_{0.11}$	$28.73_{1.04}$	$37.71_{0.17}$
	ROUGE-2 ↑	**$14.74_{0.71}$**	**$17.61_{0.06}$**	**$16.92_{0.21}$**	$18.14_{0.08}$	$9.14_{0.97}$	$13.16_{0.23}$
	ROUGE-L ↑	$28.94_{0.70}$	**$33.81_{0.47}$**	**$31.97_{0.46}$**	$33.39_{0.25}$	$22.24_{0.91}$	$29.37_{0.14}$
KSA w/o public	ROUGE-1 ↑	$36.42_{0.45}$	$34.00_{0.59}$	$40.27_{0.34}$	$35.10_{0.14}$	$28.73_{1.04}$	$37.71_{0.17}$
	ROUGE-2 ↑	$14.56_{0.36}$	$14.32_{0.22}$	$15.10_{0.21}$	$14.80_{0.32}$	$9.14_{0.97}$	$13.16_{0.23}$
	ROUGE-L ↑	$28.11_{0.39}$	$27.57_{0.31}$	$31.14_{0.46}$	$28.85_{0.12}$	$22.24_{0.91}$	$29.37_{0.14}$

Table 2: **Results for private ICL in the summarization task.** Likewise, in the QA task, the private ICL methods run for $\varepsilon = 1, 3, 8$ and ∞. All results are averaged over 5 runs with different seeds. The best results are colored for ROUGE-1, ROUGE-2, and ROUGE-L.

et al., 2016). The attacker's goal is to determine whether a target example is part of the demonstration examples used for ICL. We implement the repeat attack method (Wen et al., 2024). The intuition behind this attack is that the model tends to complete the training sentence when provided with only the first few words of the target example. The inference attack procedure works as follows: (1) The attacker **selects a target example** and attempts to determine whether it is part of the training dataset. (2) The attacker **truncates the target example** and inputs it into the model, which generates a completion. (3) The attacker **calculates the semantic similarity** between the model's completion and the target example. If the similarity exceeds a threshold τ, the attacker concludes that the target example was part of the training set.

We experiment with two different member-to-non-member ratios: (1) **Balanced scenario**: The number of member and non-member examples is equal, with a ratio of $1:1$. (2) **Unbalanced scenario**: The ratio is $1:4$, reflecting a more realistic environment. For the balanced setting, we use 40 member examples and 40 non-member examples. For the unbalanced setting, we use 40 member examples and 160 non-member examples. We use a 2-shot, 10-ensemble setting for the private model and apply SGA with $\varepsilon = [1, 3, 8]$. For the non-private baselines, we consider the non-private aggregation model ($\varepsilon = \infty$, agg) and a 20-shot model ($\varepsilon = \infty$), since the effective number of shots for the private model is $2 \times 10 = 20$. The AUROC is computed over member and non-member examples, with the results shown in Figure 3. From the results, we observe that, at all ε levels, the private model significantly mitigates the repeat attack, keeping the AUROC around or below 0.6, indicating low attack success. Notably, the non-private aggregation model also provides some defense, achieving an AUROC of 0.572 for the unbalanced setting. This suggests that non-private aggregation introduces a degree of empirical privacy, possibly due to the aggregation process reducing the influence of individual examples, even though it lacks formal differential privacy (DP) guarantees.

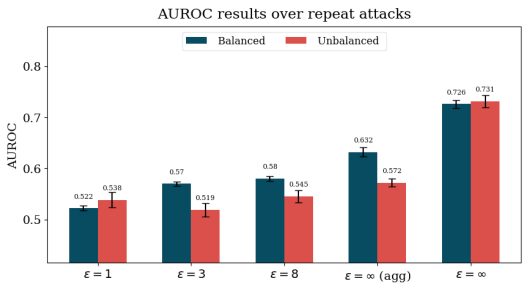


Figure 3: AUROC results for repeat attacks. "Balanced" denotes the MIA setting where the ratio between members and non-members is $1:1$, and "Unbalanced" denotes a ratio of $1:4$.

6 Further Analysis

6.1 Private ICL with OOD Public

To evaluate the performance of private ICL with an out-of-distribution (OOD) public dataset, we use the HealthCareMagic subset of ChatDoctor, which comes from a different platform, as the public dataset for the QA task. The BLEU, METEOR, and ROUGE-1 metrics are reported in Figure 4. We observe that the private ICL method with the OOD public dataset still outperforms private ICL with-

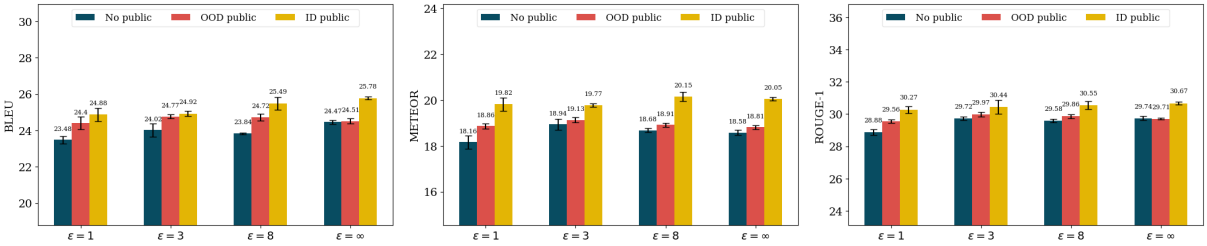


Figure 4: **Results for ICL with OOD public dataset in the question answering task.** No public data refers to private ICL without public data, OOD public to private ICL with out-of-distribution public data, and ID public to private ICL with in-distribution public data.

out public data, particularly in the strong privacy regime ($\varepsilon = 1$). However, it is outperformed by private ICL with in-distribution public data, which is expected given the alignment between the public and private datasets.

6.2 Hyperparmeter Analysis

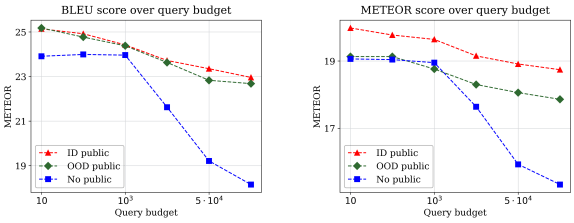


Figure 5: Comparison of BLEU and METEOR scores across query budget

Utility across varying number of query budgets We investigate the performance of private ICL under different query budgets. We fix the total privacy budget at $\varepsilon = 3$ and calculate the privacy budget for each query. We compare SGA with in-distribution (ID) public data, out-of-distribution (OOD) public data, and private-data-only baselines. The results are shown in Figure 5. Up to the query budget of 10^3, the private-data-only baseline shows comparable performance, but as the query budget increases, the performance of the private-data-only baseline quickly degrades compared to the ID public and OOD public models.

Utility across varying numbers of candidate numbers In this experiment, we use the top-k SGA model with varying numbers of candidate numbers k and privacy budget. The results are shown in Figure 6. In the non-private setting ($\varepsilon = \infty$), there are no noticeable differences across different k values. However, for $k = 2$ and $k = 8$, the model's performance degrades more quickly compared to $k = 3$ and $k = 6$. We conjecture that

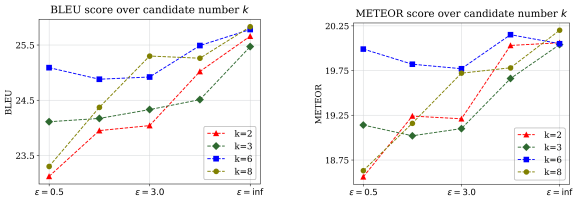


Figure 6: Comparison of BLEU and METEOR scores across candidate

a small number of candidates ($k = 2$) does not provide enough options to mitigate the effects of noisy histogram output, while a large number of candidates ($k = 8$) introduces too much confusion for the model, leading to performance degradation.

Utility across varying number of ensembles We also investigate the model's performance across different numbers of ensembles. The result is shown in Appendix D. The highest performance was observed with an ensemble size of 75. As the number of ensembles increases, performance gradually improves, but the increase is not significant due to the higher subsampling rate, which leads to a noisier histogram.

7 Conclusion

We propose a private in-context learning (ICL) framework that leverages public data. To address high dimensionality, we project the LLM output into a semantic group space. Public data is used in ensemble aggregation and selection to mitigate utility degradation from differential privacy. Experiments show an effective privacy-utility tradeoff, with notable improvements from public data. Additionally, empirical tests demonstrate strong defense against privacy and membership inference attacks across all privacy levels. Finally, we show that even out-of-distribution (OOD) public data enhances private ICL performance, offering a practical solution for real-world applications.

Limitations

One limitation of the private ICL framework is that the privacy risk accumulates over multiple queries. While we attempt to mitigate utility degradation caused by accumulated query outputs by leveraging public data, this approach may not entirely prevent privacy leakage in long-running or high-volume query settings. We also believe that this privacy accumulation issue could be alleviated by updating private demonstration examples once a query budget threshold is reached.

Another limitation is that the private ICL framework requires extensive computation due to the use of multiple ensembles. Although this ensures strong privacy guarantees, it can be computationally expensive, especially for large-scale applications. Future work could explore more computationally efficient differential privacy mechanisms or optimized ensemble methods to reduce the computational burden without sacrificing privacy.

Lastly, effective results in our framework rely on the availability of high-quality public data. This may not always be accessible in certain domains, particularly in secured areas where public data is scarce for the task. This limitation could affect the generalizability of our approach and may necessitate alternative strategies, such as synthetic data generation or domain adaptation techniques, to improve performance in such scenarios.

References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. B. McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*.

Federico Albanese, Daniel Alfredo Ciolek, and Nicolas D'Ippolito. 2023. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *ArXiv*, abs/2311.10785.

Borja Balle, Gilles Barthe, and Marco Gaboardi. 2018. Privacy amplification by subsampling: Tight analyses via couplings and divergences. *ArXiv*, abs/1807.01647.

Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. 2019. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Mark Bun and Thomas Steinke. 2016. Concentrated differential privacy: Simplifications, extensions, and lower bounds. *ArXiv*, abs/1605.02065.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey on in-context learning.

Haonan Duan, Adam Dziedzic, Nicolas Papernot, and Franziska Boenisch. 2023. Flocks of stochastic parrots: Differentially private prompt learning for large language models. *ArXiv*, abs/2305.15594.

Cynthia Dwork. 2006. Differential privacy. In *International Colloquium on Automata, Languages and Programming*.

James Flemings, Meisam Razaviyayn, and Murali Annavaram. 2024. Differentially private next-token prediction of large language models. *ArXiv*, abs/2403.15638.

Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam D. Smith. 2008. Composition attacks and auxiliary information in data privacy. In *Knowledge Discovery and Data Mining*.

Shayna Gardiner, Tania Habib, Kevin Humphreys, Masha Azizi, Frederic Mailhot, Anne Paling, Preston Thomas, and Nathan Zhang. 2024. Data anonymization for privacy-preserving large language model fine-tuning on call transcripts. In *CALDPSEUDO*.

Jennifer Gillenwater, Matthew Joseph, Andres Munoz, and Monica Ribero Diaz. 2022. A joint exponential mechanism for differentially private top-k. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7570–7582. PMLR.

Antonio A. Ginart, Laurens van der Maaten, James Y. Zou, and Chuan Guo. 2022. Submix: Practical private prediction for large-scale language models. *ArXiv*, abs/2201.00971.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Backdoor attacks for in-context learning with language models. *ArXiv*, abs/2307.14692.

Aly Kassem, Omar Mahmoud, and Sherif Saad. 2023. Preserving privacy through dememorization: An unlearning technique for mitigating memorization risks

in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4379, Singapore. Association for Computational Linguistics.

P. Mayil Vel Kumar and M. Karthikeyan. 2012. L diversity on k-anonymity with external database for improving privacy preserving data publishing. *International Journal of Computer Applications*, 54:7–13.

Haoran Li, Yulin Chen, Jinglong Luo, Yan Kang, Xiaojin Zhang, Qi Hu, Chunkit Chan, and Yangqiu Song. 2023a. Privacy in large language models: Attacks, defenses and future directions. *ArXiv*, abs/2310.10383.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2024. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *ArXiv*, abs/2406.05948.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori B. Hashimoto. 2021. Large language models can be strong differentially private learners. *ArXiv*, abs/2110.05679.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).

Jimit Majmudar, Christophe Dupuy, Charith Peris, Sami Smaili, Rahul Gupta, and Richard S. Zemel. 2022. Differentially private decoding in large language models. *ArXiv*, abs/2205.13621.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *ArXiv*, abs/2202.12837.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. 2023. Effectively using public data in privacy preserving machine learning. In *International Conference on Machine Learning*.

Anthia Papadopoulou, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. Neural text sanitization with explicit measures of privacy risk. In *AACL*.

Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Úlfar Erlingsson. 2018. Scalable private learning with pate. *ArXiv*, abs/1802.08908.

R. Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2016. Membership inference attacks against machine learning models. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Robin Staab, Mark Vero, Mislav Balunovi'c, and Martin T. Vechev. 2024. Large language models are advanced anonymizers. *ArXiv*, abs/2402.13846.

Xinyu Tang, Richard Shin, Huseyin A. Inan, Andre Manoel, FatemehSadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. 2023. Privacy-preserving in-context learning with differentially private few-shot generation. *ArXiv*, abs/2309.11765.

Jun Wang and Zhi-Hua Zhou. 2020. Differentially private learning with small public data. In *AAAI Conference on Artificial Intelligence*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed Huai hsin Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *ArXiv*, abs/2206.07682.

Rui Wen, Zheng Li, Michael Backes, and Yang Zhang. 2024. Membership inference attacks against in-context learning. *ArXiv*, abs/2409.01380.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. 2024. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*.

Da Yu, Sivakanth Gopi, Janardhan Kulkarni, Zi-Han Lin, Saurabh Naik, Tomasz L. Religa, Jian Yin, and Huishuai Zhang. 2023. Selective pre-training for private fine-tuning. *ArXiv*, abs/2305.13865.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially private fine-tuning of language models. *ArXiv*, abs/2110.06500.

A Details of Differential Privacy

A.1 JointEM algorithm

For the differential private top-k, we use JointEM from (Gillenwater et al., 2022). The JointEM algorithm is one of the exponential methods that uses count difference as the utility scores. First, to sample from the utility score, the line $7-22$ estimates the number of sequences that has utility score $\tilde{U}_{ij}$. Here, $\tilde{m}(\tilde{U}_{ij})$ denotes the number of sequences that has utlity $\tilde{U}_{ij}$. After calculating $\tilde{m}$ values, in the line $23-28$, first allocate index j for s_i and sample remaining item $k-1$ items by sampling at random from $\left\{t_l(\tilde{U}_{ij})\right\} \setminus \{j, s_1, \ldots, s_{l-1}\}$.

A.2 Privacy Accounting

As privacy accounting involves composition, we need the following theorems for tight accounting.

Theorem 1 (Balle et al. (2018)) *Let* `Uniform` *denotes sampling m elements from n data points without replacement Let $\mathcal{M}' = \mathcal{M} \circ$* `Uniform`*. For any $\varepsilon \geq 0$ we have $\delta_{\mathcal{M}'}(\varepsilon') \leq (m/n)\delta_{\mathcal{M}}(\varepsilon)$, where $\varepsilon' = \log(1 + (m/n)(e^{\varepsilon} - 1))$.*

Theorem 2 (DP to RDP Bun and Steinke (2016)) *The exponential mechanism is ε-DP and $(\alpha, \varepsilon_{\mathsf{EM}}(\alpha))$-RDP, where $\varepsilon_{\mathsf{EM}}(\alpha)$ is defined as*

$$\min\left(\frac{\alpha}{2}\varepsilon^2, \frac{1}{\alpha-1}\log\left(\frac{\sinh(\alpha\varepsilon)-\sinh((\alpha-1)\varepsilon)}{\sinh(\varepsilon)}\right)\right)$$

Theorem 3 (RDP composition) *If each mechanism $(\mathcal{M}_i)_{i=1}^k$ is (α, ε)-RDP, then the composition $A_k \circ A_{k-1} \circ \cdots \circ A_1$ is $(\alpha, k\varepsilon)$-RDP.*

Theorem 4 (RDP to approximate DP Balle et al. (2019))

If a mechanism $\mathcal{M}$ is (α, ρ)-RDP then it is $(\rho + \log((\alpha-1)/\alpha) - (\log\delta + \log\alpha)/(\alpha-1), \delta)$-DP for any $0 < \delta < 1$.

First, express privacy loss jointEM in terms of RDP using theorem 2. Then the calculate privacy loss composed jointEM by calling `compose_subsampled_EM` API using `AutoDP` package. Then, the converting RDP to approx DP using theorem 4.

Algorithm 3 `JointEM` (Gillenwater et al., 2022)

Require: Vectors of item counts $c_1, \ldots, c_d$, number of items to estimate k, privacy parameter ε

1: Sort and relabel items so $c_1 \geq c_2 \geq \cdots \geq c_d$
2: Construct $k \times d$ matrix $\tilde{U}$ by
$$\tilde{U}_{ij} = -(c_i - c_j) - \frac{d(k-i)+j}{2dk}$$
3: Sort $\tilde{U}$ in decreasing order to get $\tilde{U}_{(1)}, \ldots, \tilde{U}_{(dk)}$, storing the (row,column) of each $\tilde{U}_{(a)}$ as $(r(a), c(a))$
4: Initialize $n_1, \ldots, n_k \leftarrow 0$
5: Initialize set of non-zero n_i, $N \leftarrow \emptyset$
6: Initialize $b \leftarrow 0$
`// Handle` $\tilde{m}$ `that are zero`
7: **for** $a = 1, \ldots, dk$ **do**
8: $\quad n_{r(a)} \leftarrow c(a) - (r(a) - 1)$
9: $\quad N \leftarrow N \cup \{r(a)\}$
10: $\quad$ **if** $|N| = k$ **then**
11: $\quad\quad$ break
12: $\quad$ **end if**
13: $\quad$ Set $\tilde{m}(\tilde{U}_{(a)}) \leftarrow 0$, and set $b \leftarrow a$
14: **end for**
`// Handle` $\tilde{m}$ `that are non-zero`
15: Set $p \leftarrow \Pi_{r \in [k]} n_r$
16: Compute $\tilde{m}(\tilde{U}_{(b+1)}) \leftarrow p/n_{r(a)}$
17: **for** $a = b+2, \ldots, dk$ **do**
18: $\quad p \leftarrow p/n_{r(a)}$
19: $\quad$ Compute $\tilde{m}(\tilde{U}_{(a)}) \leftarrow p$
20: $\quad$ Update $n_{r(a)} \leftarrow n_{r(a)} + 1$
21: $\quad$ Update $p \leftarrow p \cdot n_{r(a)}$
22: **end for**
23: Sample a Utility $\tilde{U}_{ij}$ from:
$$\mathbb{P}[\tilde{U}_{ij}] \propto \tilde{m}(\tilde{U}_{ij}) \exp\frac{\varepsilon\left[\tilde{U}_{ij}\right]}{2}$$
24: Initialize size-k output vector s with $s_i \leftarrow j$
25: **for** $l = 1, 2, \ldots, i-1, i+1, \ldots, k$ **do**
26: $\quad$ Compute $t_l(\tilde{U}_{ij})$ by iterating through row l of $\tilde{U}$
27: $\quad$ Sample s_l uniformly from $\left\{t_l(\tilde{U}_{ij})\right\} \setminus \{j, s_1, s_2, \ldots, s_{l-1}\}$
28: **end for**
29: **return** Vector of item indices s

B Examples of repeat attack

Examples of successful and failed attacks. The blue part is provided to the model for completion.

Repeat attack examples

Sucessful Attack Examples

Model Completion: hi, my doughter(3 yrs) is having cold cough and wheezing very frequently...and she'll get asthaline and sometimes budecort nebulization and antibiotic....this time (she had mild infection) she got clarithromycin antibiotic 6ml twice a day for 5 days....but she doesn't like the taste and spills all the medicine somehow I managed to give her for 2 days with spilling and all...so I stopped for 1 day.....now my question is can I give her azithromycin because she takes that easily ...will that be effective?

Target Question: hi, my doughter(3 yrs) is having cold cough and whizzing very frequently...and she,ll get asthaline and sometimes budecort nabulization and antibiotic....this time(she had mild infection) she got clarithromycin antibiotic 6ml twice a day for 5 days....but she does,t like the taste and spilling all the medicine somehow i managed to give her for 2 days with spilling and all...so i stopped for 1 day.....now my question is can i give her azithromycin bcoz she takes that easily ...will that be effetive

Score: 0.983

Model Completion: Sir,Iam suffering from shoulder neck pain for a pretty long time. Recently, I am having pain in the right side of my brain above my head followed by pain in the eye. My MRI report says T̈iny T2 hyperintense foci in bilateral frontoparietal subcortical white matter suggestive of ischemic foci.Ẅhat does this mean? Is any treatment required, and if so, what is your advice?

Target Question: Sir,Iam suffering from shoulder neck pain for a pretty long time.Recently i am having pain in the right side of of brain above head followed by pain in the eye.MRI {P+C) report says Tiny T2 hyper intensefoci in bilateral frontoparietal subcortical white matter suggestive of ischemic foci.sir,what does this mean .any treatment required and if so what is your advice?

Score: 0.981

Failed Attack Examples

Model Completion: Hello. For a week now I have been experiencing frequent urination and a burning sensation when I pee. I also feel pressure in my lower abdomen. I have not had any urinary tract infections before. What could be causing these symptoms? Thank you.

Target Question: Hello. For a week now I have been having productive cough. During the first few days, I was feeling lethargic and I was coughing a lot. Then the lethargy went away and I coughed less (about 3 days in). But then I started coughing up small amounts of blood with sputum, about once a day. Then the cough almost went away and I started having a runny nose which turned into congestion and now something like a sinus infection. I still get the blood in sputum about once a day. Its been 4 days now.

Score: 0.764

C Dataset example

C.1 ChatDoctor ICLineq

ChatDoctor ICLiniq

Question: Hello doctor,I got my wisdom tooth extracted a few weeks ago, and I am concerned about a piece of red tissue growth in my socket. I am not sure if it is just a new forming tissue. Please clarify.

Answer: Hello. 1. It looks like a piece of surrounding gum or gingival tissue remaining after extraction. 2. I wanted to know when did you notice this for the first time? Was it not there right after extraction, or did you not notice? 3. From photograph/your attachment, it does not look like anything major. (attachment removed to protect patients identity). 4. If you feel pain or sensitivity in it, I would suggest you consult your dentist. He will remove it from the site, and it will not interfere any more. 5. Just keep a notice of whether it is increasing or decreasing in size or any color changes. 6. Avoid very hot and spicy food at least from that side for some time. 7. Report me if you notice any changes. Follow up after a week.

C.2 ChatDoctor HealthcareMagic (OOD)

ChatDoctor HealthcareMagic

Question: Masses full thickness tear of the whole subscapularis tendon with associated atrophy of the musel belly suggestive of a chronic tear ! complete full thickness tear of the whole subscapularis tendon with tendon retraction and a complete tear of the long head biceps tendon with tendon retraction! and large joint effusion

Answer: Hello welcome to the Chat Doctor You are having complete muscle tear and so joint effusion as well. CT scan report should be done to look for grade of muscle tendon injury. Intraluminal steroid can be given to relieve inflammation If age and clinical condition permit than surgical treatment is needed To Hope this will help you Take care Consult orthopedic surgeon for examination and discuss your case

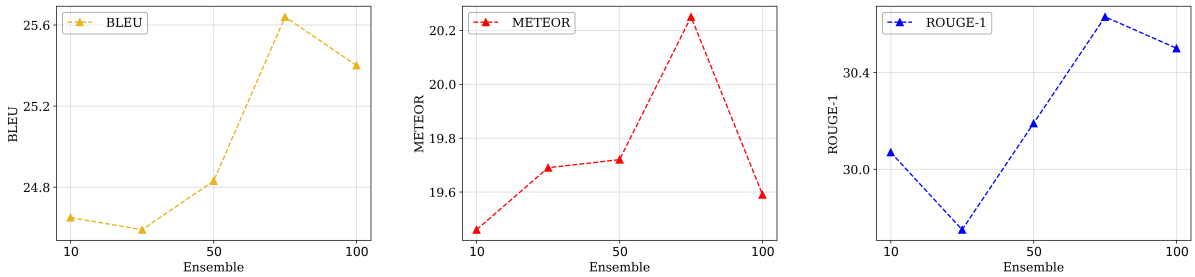
C.3 SAMsum

SAMsum

Dialogue: Dialogue: "Liam: <file_other>
Liam: have you read this one?
Julia: wow, I didn't even know that it existed
Julia: thanks! The last one I read was Die again
Liam: this one is like from 2014?
Julia: yep
Liam: I've heard that city library is organizing a meeting with Tess
Julia: really? That would be really awesome
Liam: just check it out on their facebook"

Reference Summary: The last one Julia read was Die again from 2014. There's going to be a meeting with Tess organized by the city library.

D Experiment with varying ensemble numbers



(a) BLEU score over candidate number k

(b) METEOR score over candidate number k

(c) ROUGE-1 score over candidate number k

Figure 7: Comparison of performance scores across varying ensemble numbers

E Hyperparameters

Dataset	τ	k	JointEM ε' for $\varepsilon = [1, 3, 8]$	δ
ChatDoctor	0.95	$(6, 40)$	$[0.23, 0.63, 1.32]$	$2.56 \cdot 10^{-4}$
SAMsum	0.92	$(6, 10)$	$[0.53, 1.25, 2.22]$	$6.8 \cdot 10^{-5}$

Table 3: Hyperparmeter setting. k denotes the candidate number for SGA and keyword candidate number for KSA.

F Prompt construction for candidate selection

F.1 Candidate selection prompt for SGA (QA)

ChatDoctor iclinq

Instruction: You are a doctor. Please answer the medical questions based on the patient's description

<Public Demonstration Example>

<Question>

Pick the most accurate answer for the question with the following answer candidates ranked by their frequency from high to low: **[<Candidates>]**

The answer is:

F.2 Candidate selection prompt for SGA (Summarization)

SAMsum

<Public Demonstration Example>

<Dialogue>

Pick the most accurate summary for the dialogue with the following summary suggestions: **[<Candidates>]**

The summary is:

F.3 Prompt construction for KSA (QA)

ChatDoctor iclinq

Instruction: You are a doctor. Please answer the medical questions based on the patient's description

<Public Demonstration Example>

<Question>

Answer the above question with the following word suggestions ranked by their frequency from high to low: **[<Suggestions>]**

The answer is:

F.4 Prompt construction for KSA (Summarization)

ChatDoctor iclinq

<Public Demonstration Example>

<Dialogue>

Summarize the above dialogue with the following word suggestions ranked by their frequency from high to low:: **[<Candidates>]**

The summary is:

G SGA output examples

G.1 ChatDoctor iclinq

ChatDoctor iclinq

Question: Hello,I am a 27 year old female. I am feeling my heartbeat, breath, and pulse rate to be really abnormal since the past three weeks. It feels like my heart is pounding so bad. And then, I feel a difficulty to breathe, like I have run or jogged more than my capacity, even if I am sitting or lying. Sometimes, I feel a lot of pain in my lower abdomen area and on the left side of my belly. Please suggest me some remedy.

($\varepsilon = 1$) **Answer:** Hello. It is important to get these symptoms evaluated by a healthcare professional to determine the underlying cause. The symptoms you are experiencing could be related to various conditions such as anxiety, heart issues, respiratory problems, or gastrointestinal issues. Given the significant symptoms you are experiencing, it is recommended that you seek urgent medical attention. Please visit your doctor or the nearest emergency department for a thorough evaluation and appropriate management. It is crucial not to ignore these symptoms, especially if they are causing you significant distress and discomfort.

($\varepsilon = 3$) **Answer:** Hello. It is important to get these symptoms evaluated by a healthcare professional to determine the underlying cause. The symptoms you are experiencing could be related to various conditions such as anxiety, heart issues, respiratory problems, or gastrointestinal issues. Given the significant symptoms you are experiencing, it is recommended that you seek urgent medical attention. Please visit your doctor or the nearest emergency department for a thorough evaluation and appropriate management. It is crucial not to ignore these symptoms, especially if they are causing you significant distress and discomfort.

($\varepsilon = 8$) **Answer:** Hello, Welcome to Chat Doctor forum. Your symptoms of feeling abnormal heartbeats, pounding sensation, difficulty in breathing, and pain in the lower abdomen area should not be ignored. It is important to seek medical attention promptly for a proper evaluation of your symptoms. These symptoms could be due to various reasons such as anxiety, stress, hormonal imbalances, heart conditions, or other medical issues. It is recommended that you schedule an appointment with a healthcare provider for a physical examination, which may include tests such as an electrocardiogram (ECG), blood tests, and possibly a chest X-ray to determine the cause of your symptoms. In the meantime, try to practice relaxation techniques, deep breathing exercises, and avoid caffeine and other stimulants. Take care.

Reference: Hello, Welcome to Chat Doctor forum. I went through your post. Your predominant complaints are palpitations and lower abdomen pain. Any person who notices a change in their heartbeats should go for ECG (electrocardiogram) to rule out arrhythmia. The most common causes of a one-sided pain in the lower abdomen are kidney stone or urinary tract infection. These are only the most common causes. However, there are more than 10 causes of the same. A USG (ultrasonogram) of the abdomen with a urine examination will help you come to a conclusion. Consult a physician with above investigations so he can help you. Regards. For more information consult an internal medicine physician online

G.2 SAMsum

SAMsum

Dialogue: Jake: What are your plans for the day?
Olivia: I haven't really got anything planned. There are some things I should do which I haven't looked at for a while
Jake: Like what?
Olivia: I ought to do my accounts. At least get all the incoming and outgoing invoices sorted by months.
Jake: Tax filing is a long way off.
Olivia: I know, but doing it all at the last minute is a dreadful headache.
Jake: Anything else?
Olivia: I also have a few videos I need to upload to YouTube.
Jake: Already ready or you still need to do post-production?
Olivia: I don't really do much by way of post-production. I'm not exactly Steven Spielberg.
Jake: I do, but I dont always have time to do much.
Olivia: If I put in any music I always get a copyright strike and someone else is taking the ad revenue
Jake: Yeah. And some of those claims are purely speculative. One time I put on one of my own pieces and it was challenged by Sony Music
Olivia: No way!
Jake: It's true. On the form you fill in to counter the claim I said if Sony want to make me one of their signed artists, that's fine. Otherwise to get they paws off of my music."

($\varepsilon = 1$) **Summary:** Olivia and Jake talk about their plans for the day. Olivia has some videos that she needs to upload to YouTube and Jake wants to share his own music.

($\varepsilon = 3$) **Summary:** Olivia doesn't have any plans for the day. Olivia does her accounts. Olivia uploads videos to YouTube. Olivia also does post-production on her videos. Olivia puts in music in her videos but always get a copyright strike. Olivia's videos sometimes get claims by Sony Music. Olivia put her own piece of music once on YouTube and it was challenged by Sony Music.

($\varepsilon = 8$) **Summary:** Olivia says she has some things to do, such as sorting incoming and outgoing invoices by month. Jake says he does videos for Youtube. He has not got much time for post-production, though.

Reference: Olivia has to sort out her accounts and upload a few videos on YouTube. Jake is complaining that Sony Music tried to appropriate his own music.

H KSA output examples

H.1 ChatDoctor iclineq

ChatDoctor iclinq

Question: Hello,I am a 27 year old female. I am feeling my heartbeat, breath, and pulse rate to be really abnormal since the past three weeks. It feels like my heart is pounding so bad. And then, I feel a difficulty to breathe, like I have run or jogged more than my capacity, even if I am sitting or lying. Sometimes, I feel a lot of pain in my lower abdomen area and on the left side of my belly. Please suggest me some remedy.

($\varepsilon = 1$) **Answer:** Please suggest a remedy for your symptoms. It is crucial to assess your symptoms, especially concerning your heartbeats and difficulty breathing. It is best to seek evaluation at a clinic or medical facility to identify the nature of your symptoms and ensure safe and appropriate treatment. Do not hesitate to seek medical attention, as your symptoms could potentially be due to a cardiac disorder or other serious condition. Managing your symptoms and relieving any pain is important to prevent any potential worsening of your condition. It is also important to separate any possible infections or embolism as a cause of your symptoms. It is possible that your symptoms may be related to nicotine use, so it is important to address this as well.

($\varepsilon = 3$) **Answer:** Please seek medical attention immediately. It is important to identify the causes of your symptoms, which may be associated with palpitations, breathing difficulties, and pain in your abdomen. A thorough evaluation, including an ECG, echocardiogram, and possibly other tests, is recommended to determine the severity and appropriate treatment. Do not ignore these symptoms, especially if they are severe or associated with fainting or other concerning signs. It is important to prioritize your health and well-being by seeking help from a doctor or healthcare center.Ignoring these symptoms can lead to serious health issues, so please do not delay in seeking medical help.

($\varepsilon = 8$) **Answer:** It is essential to note that abnormal heart beating, difficulty in breath, and pain in the abdomen area are concerning signs that may indicate a heart-related issue. It is advisable to seek professional attention and start proper treatment. Keeping a separate note of the frequency and various techniques to understand and recommend the appropriate exercises to avoid strain is advisable. It is important to avoid strenuous activities and seek detailed advice from a healthcare professional. If experiencing dizziness or lightheadedness, it is important to seek medical attention promptly.

Reference: Hello, Welcome to Chat Doctor forum. I went through your post. Your predominant complaints are palpitations and lower abdomen pain. Any person who notices a change in their heartbeats should go for ECG (electrocardiogram) to rule out arrhythmia. The most common causes of a one-sided pain in the lower abdomen are kidney stone or urinary tract infection. These are only the most common causes. However, there are more than 10 causes of the same. A USG (ultrasonogram) of the abdomen with a urine examination will help you come to a conclusion. Consult a physician with above investigations so he can help you. Regards. For more information consult an internal medicine physician online

H.2 SAMsum

SAMsum

Dialogue: Jake: What are your plans for the day?
Olivia: I haven't really got anything planned. There are some things I should do which I haven't looked at for a while
Jake: Like what?
Olivia: I ought to do my accounts. At least get all the incoming and outgoing invoices sorted by months.
Jake: Tax filing is a long way off.
Olivia: I know, but doing it all at the last minute is a dreadful headache.
Jake: Anything else?
Olivia: I also have a few videos I need to upload to YouTube.
Jake: Already ready or you still need to do post-production?
Olivia: I don't really do much by way of post-production. I'm not exactly Steven Spielberg.
Jake: I do, but I dont always have time to do much.
Olivia: If I put in any music I always get a copyright strike and someone else is taking the ad revenue
Jake: Yeah. And some of those claims are purely speculative. One time I put on one of my own pieces and it was challenged by Sony Music
Olivia: No way!
Jake: It's true. On the form you fill in to counter the claim I said if Sony want to make me one of their signed artists, that's fine. Otherwise to get they paws off of my music."

($\varepsilon = 1$) **Summary:** Olivia has a lot of things to do, but she doesn't have time to do them. She has to do her accounts, upload videos to YouTube, and edit her videos. She doesn't have time to do post-production. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want to get a copyright strike. She doesn't want.

($\varepsilon = 3$) **Summary:** Olivia and Jake are talking about their plans for the day. Olivia has some things she needs to do, like sorting out her accounts and uploading some videos to YouTube. Jake says he does some post-production work on his videos, but he doesn't always have time to do much. Olivia says she doesn't do much post-production work because she always gets a copyright strike and someone else is taking the ad revenue. Jake says he has had a copyright strike from Sony Music on one of his own pieces of.

($\varepsilon = 8$) **Summary:** Olivia and Jake are talking about their plans for the day. Olivia has some things she needs to do, like sorting out her accounts and uploading some videos to YouTube. Jake says he does some post-production work on his videos, but he doesn't always have time to do much. Olivia says she doesn't do much post-production work because she always gets a copyright strike and someone else is taking the ad revenue. Jake says he has had a copyright strike from Sony Music on one of his own pieces of.

Reference: Olivia has to sort out her accounts and upload a few videos on YouTube. Jake is complaining that Sony Music tried to appropriate his own music.

I Information About Use Of AI Assistants

In the writing process, we use an AI assistant to fix grammar and typos in the manuscript.